Executive Summary: Statistical Methods for 2021 Coverage Determinations under Voting Rights Act Section 203(b)

by Carolina Franco and Eric Slud, CSRM

October 25, 2021

According to Section 203(b) of the *Voting Rights Act of 1965* beginning in 1975, and as amended in 1982 and 2006, states and political subdivisions must in certain circumstances make voting materials available in languages other than English. These circumstances are defined in Section 203(b) in terms of specific determinations involving the sizes and proportions of designated population subgroups as measured by the decennial census and the most current American Community Survey (ACS). Section 203(b) as amended prescribes that the Director of the Census Bureau shall make these determinations every 5 years, based on the most current population estimates derived from the ACS along with relevant census data. The 2021 determinations to be released in December 2021 are based solely on 2015-2019 5-year ACS data. The decennial 2020 Census data were not used. This decision was made as a response to concerns about the timing of the decennial counts, which was affected by the COVID-19 Pandemic.

For the determinations, estimates are needed at various levels of geographic aggregation. These levels of geography include states, jurisdictions, American Indian Areas (AIAs), and Alaskan Native Regional Corporations (ANRCs). The nation is partitioned into roughly 8000 Jurisdictions (7859 in ACS 2015-2019 5-year data containing at least one voting-age respondent), which are Counties in most states and Minor Civil Divisions (MCDs) in the other states. Other geographic domains relevant to provisions of Section 203(b) are the American Indian Areas (AIAs), of which there are 568 with ACS respondents in 2015-2019, as well as 12 Alaska Native Regional Corporations (ANRCs). All 12 ANRCs had at least one person in the sample.

For purposes of Section 203(b), only the population of voting age (18 or over) persons is relevant. The law categorizes voting age persons according to Citizenship, Limited English Proficiency (LEP) and Illiteracy. The classifications by voting age, Citizenship and Illiteracy, are each defined by the answer to a single ACS question, and LEP is defined through the answers to two ACS questions. People self-identify (in the Census or ACS) as belonging to one or more of 6 distinct racial groups, (each containing several specific races) and 1 ethnic classification

that are then used to define 73 'Language Minority Groups' (LMGs) for purposes of Section 203(b). Of the LMGs, 21 are Asian, 51 are American Indian or Alaska Native (AIAN), and one is Hispanic.

Section 203(b) prescribes generally that states and political subdivisions are required to provide voting materials in a language other than English for members of a LMG according to the following rules:

(i) A state must do so if the illiteracy rate among LEP members of the LMG in the state exceeds the national rate of illiteracy among citizens, **and** the number of LEP persons in the LMG is greater than 5% of the total number of voting-age Citizens in the state.

(ii) A jurisdiction must do so if the illiteracy rate among LEP persons in the LMG and jurisdiction exceeds the national rate of citizen illiteracy **and** the number of LEP persons in the jurisdiction and LMG is greater than either 10,000 or 5% of the total CIT population of the jurisdiction.

(iii) An American Indian Area (AIA) or Alaskan Native Regional Corporation (ANRC) *and all jurisdictions containing any part of it* must do so for an AIAN LMG if the illiteracy rate among LEP AIAN persons of the LMG in the AIA/ANRC exceeds the national rate of citizen illiteracy **and** the number of LEP AIAN persons in the AIA/ANRC and LMG is greater than 5% of the total voting-age citizen AIAN population of the AIA/ANRC.

Special tabulations of weighted survey estimates of state, jurisdiction, AIA, and ANRC voting-age populations cross-classified by citizenship, limited English proficiency, illiteracy, and LMG are available from ACS 5-year data. These tabulations could be used to create direct estimates of all of the ingredients of the 'triggering' criteria (i)-(iii) for determinations. However, the counts of ACS sampled voting-age persons by jurisdiction and LMG on which these weighted sums would be based are often quite small, and the variability (standard errors) of the direct estimates are often quite large compared to the estimates themselves. Moreover, the standard errors estimated by current ACS methodology are also very unreliable for population counts in such small domains.

For that reason, starting in 2011 and again in 2016 and 2021, statistical research on the estimation methodology driving the Section 203 determinations has been primarily directed toward model-based, 'small area estimation'. Small area estimation is devoted to enhancing the precision of estimation through the formulation of models for multiple small areas which 'borrow strength' from one another through

shared statistical parameters, and through the use of auxiliary information.

The main idea of this approach is that jurisdictions within the same LMG may behave similarly with respect to the characteristics of interest across different geographies, and with respect to covariates.

The domains used for the small area estimation models are Jurisdictions for each of the LMGs, and AIAs or ANRCs for each of the AIAN LMGs. Statistical models are fitted separately for the different LMGs and types of geography. In addition, the complexity of the model used for a particular LMG and type of geography depends on how many distinct geographic units have ACS respondents for that LMG. This is necessary as the ACS sample for some LMGs and geography types contains thousands of people, while for other LMGs and types of geography the ACS sample may contain only a single person.

The general form of model chosen for the 2021 statistical estimation is a Multinomial Logit Normal (MLN) Model, formulated for the nested decreasing subpopulations of voting age persons (VOT), voting age citizens (CIT), voting age citizens who are limited English proficient (LEP), and illiterate limited English Proficient voting age persons (ILL). The MLN model is a random-effects generalization of logistic regression, in which the proportions of CIT persons within VOT, and similarly LEP within CIT and ILL within LEP, are modeled using a logit transformation and random intercepts, as well as predictive covariates.

The covariates used in modeling were computed from the same ACS dataset as the response variable but at higher levels of aggregation. One set of covariates was defined as the higher-geography-level LMG proportion of CIT within VOT, LEP within CIT, and ILL within LEP, for the portion of the State complementary to a Jurisdiction it contains, or the portion of the whole AIAN LMG complementary to an AIA. All other covariates were defined at the level of the geographic unit, without regard to LMG. One such covariate, in all geography types, was the proportion of people speaking a language other than English in the home. Covariates used in Asian and Hispanic LMGs include the proportion of Foreign-born, the average years in US for the foreign-born, and the proportions in coarse age-groups. Covariates used in various AIAN LMGs include the proportions of high-school graduates, of white nonhispanic people, and of people in poverty.

In the most detailed form of the model, the random intercepts for the CIT, LEP and ILL sub-models were jointly normally distributed and dependent. In less data-rich LMGs, the random intercepts were assumed independent. In LMGs with still less data, models of this form with reduced sets of covariates – or with none

at all – were fitted. In smaller (AIAN) LMGs in which the submodel CIT rates were uniformly close to 1, or in which the LEP or ILL rates were uniformly close to 0, an even simpler form of model was fitted. This was a common-intercept beta-binomial model with no covariates or random effects, which amounts to fitting a single rate on the pooled LMG data.

The models chosen have been explored extensively in practice data analyses using ACS 2014-2018 5-year data in the same way that the model was ultimately employed on ACS 2015-2019 5-year data. The model has been assessed against the direct domain population estimators obtained from the ACS and to those obtained by a Dirichlet multinomial model closely related to the model used in producing 2016 determinations. Model diagnostics were used in selecting covariates for the models and in assessing the suitability of the final models chosen. These analyses will be elaborated in the technical documentation. Uncertainty estimation was based on either Markov Chain Monte Carlo computation of posterior variances or a Successive Differences Replication method applied to the modeled estimates, depending on the complexity of the model.

Though all counts and proportions for Jurisdictions, AIAs, and ANRCs cross-classified by LMGs were modeled, direct ACS estimates were used for quantities at higher levels of aggregation, such as state-level estimates, or estimates by Jurisdiction that are not cross-classified with LMG. In addition, direct ACS estimators of voting age persons by LMG and geography were used to translate proportion estimates from the models into corresponding population counts. The uncertainty of these direct ACS estimates of voting age person counts was taken into account when computing the variances of the corresponding ILL, LEP, and CIT counts.

Most of the determinations are the same using the model as those that would have been obtained via the direct estimators, but there are some cases in which the model would give a determination where the direct estimates would not, and vice-versa. The direct estimators can be quite volatile and unreliable for domains with small sample sizes, and there are many such domains for LMGs in ANRCs, AIAs, and even Jurisdictions. The model predictions are more stable and result in a substantial decrease in estimates with large Coefficients of Variation (CVs, e.g. > 0.6), and in large overall reductions in Margins of Errors (MOEs). More detailed comparisons of the CVs and MOEs will be included in the technical report.

There are several ways in which the modeling approach adopted in 2021 differed from that used in 2016. First was the overall class of models chosen, Multinomial Logit Normal in place of Dirichlet-multinomial. The MLN model has more param-

eters (because of the general dependence among random intercepts), which were reduced in less data-rich LMGs by assuming the three CIT, LEP and ILL random intercepts to be independent. Second was the choice to model all predictions in LMG by geography domains (below the level of States), no matter how data-sparse. A third distinction in modeling arose because in 2016, LMG by geography domains with sample smaller than 5 (or in some cases 3) were not used in fitting LMG-level model parameters, while all LMG by geography domains with respondents were used in 2021. Finally, the variances of estimated totals and proportions were estimated in 2021 by a combination of Bayesian posterior variances (in the models for Jurisdictions in the largest LMGs) and replicate-weights based on repeated calculation of estimates with alternate weights, while in 2016 the variance estimates were calculated by a hybrid method combining parametric bootstrap and replicate weights.